Chapter 5: Complex Regressors

1 Factors

Factors allow the inclusion of qualitative or categorical predictors in the mean function of a multiple

linear regression model. Factors can have two levels, such as male and female, treated or untreated, and so on, or they can have more than two levels, such as eye color, location, or type of business.

1.1 One-factor models

Consider the United Nations data described in Section 3.1, which is an observational study of all $n = 199$ localities. The factor we use is called group,

which classified the countries into three categories: africa, oecd, and other. With no predictors beyond group, the model we fit returns estimated mean values for lifeExpF for each level of group. This is called a one-factor design or a one-way design.

Factor predictors can be included in a multiple linear regression using *dummy variables*. Since group has $d = 3$ levels, the jth dummy variable U_j for factor, $j = 1, 2, \ldots, d$ has ith value u_{ij},

for $i = 1, 2, \ldots, n$, given by

$$u_{ij} = \begin{cases} 1, & \text{if group}_i = j\text{th category of group,,} \\ 0, & \text{otherwise.} \end{cases}$$

To avoid the collinearity $U_1 + U_2 + U_3 = 1$, provided the intercept term is included, one of the dummy variables need to be dropped. For example, dropping U_1, we will have

$$E(lifeExpF|group) = \beta_0 + \beta_2 U_2 + \beta_3 U_3.$$

Table 1 summarizes the fit of the one-way model.

The means for the three groups are

$$\begin{aligned}
\hat{E}(lifeExpF|group = oecd) &= \hat{\beta}_0 + \hat{\beta}_2 0 + \hat{\beta}_3 0 \\
&= 82.45, \\
\hat{E}(lifeExpF|group = other) &= \hat{\beta}_0 + \hat{\beta}_2 1 + \hat{\beta}_3 0 \\
&= 82.45 - 7.12, \\
\hat{E}(lifeExpF|group = oecd) &= \hat{\beta}_0 + \hat{\beta}_2 0 + \hat{\beta}_3 1 \\
&= 82.45 - 22.67.
\end{aligned}$$

Table 1: Regression summary for a factor model on UN11 data.

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           82.446      1.128  73.095  < 2e-16 ***
factor(group)other    -7.120      1.271  -5.602  7.1e-08 ***
factor(group)africa  -22.674      1.420 -15.968  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 6.28 on 196 degrees of freedom
Multiple R-squared:  0.6191,Adjusted R-squared:  0.6152
F-statistic: 159.3 on 2 and 196 DF,  p-value: < 2.2e-16
```

1.2 Comparison of Level Means

To compare the means pairwise in general requires computing the standard error of the difference between each pair of means. For example, to compare the means between the groups of other and africa, i.e., $\beta_2 - \beta_3$, we let $\boldsymbol{a} = (0, 1, -1, 0)'$ so $l = \boldsymbol{a}'\boldsymbol{\beta} = \beta_2 - \beta_3$. Therefore

$$
\begin{aligned}
se(\hat{l}|\boldsymbol{X}) &= \hat{\sigma}\sqrt{\boldsymbol{a}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{a}} \\
&= \hat{\sigma}\sqrt{c_{22} + c_{33} - 2c_{23}},
\end{aligned}
$$

where c_{ij} is the (i, j)th element of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. In R, the function *vcov* applied to a regression model returns $\hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

Table 2: Pairwise comparisons of level means for Group

Comparison	Estimate	SE	t-value	p-value
oecd-other	7.12	1.27	5.60	0.000
oecd-africa	22.67	1.42	15.97	0.000
other-africa	15.55	1.04	14.92	0.000

1.3 Adding a Continuous Predictor

As an additional predictor, we add log(ppgdp), the per person gross domestic product in the country, as a measure of relative wealth. The model is generally parameterized using main effects and inter-

actions, as

$$E(lifeExpF|log(ppgdp) = x, group) = \beta_0$$
$$+ \beta_{02}U_2 + \beta_{03}U_3 + \beta_1 x + \beta_{12}U_{2x} + \beta_{13}U_{3x}.$$

In R, it can be fitted by

$$lifeExpF \sim group + log(ppgdp) + group : log(ppgdp)$$

Table 3: Regression summary for UN11 data: a model with interaction terms

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                59.2137    15.2203   3.890 0.000138 ***
groupother                -11.1731    15.5948  -0.716 0.474572
groupafrica               -22.9848    15.7838  -1.456 0.146954
log(ppgdp)                  2.2425     1.4664   1.529 0.127844
groupother:log(ppgdp)       0.9294     1.5177   0.612 0.540986
groupafrica:log(ppgdp)      1.0950     1.5785   0.694 0.488703
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 5.129 on 193 degrees of freedom
Multiple R-squared:  0.7498,Adjusted R-squared:  0.7433
F-statistic: 115.7 on 5 and 193 DF,  p-value: < 2.2e-16
```

Examination of Table 3 suggests that while intercepts might differ for the three levels of group, the slops may be equal. This suggests fitting a model that allows each group to have its own intercept, but all groups have the same slope:

$$E(lifeExpF|log(ppgdp) = x, group) = \beta_0 + \beta_{02}U_2 + \beta_{03}U_3 + \beta_1 x,$$

which is called the main effects model. The regression is summarized in Table 4.

2 Many factors

Increasing the number of factors or the number of continuous predictors in a mean function can add considerably to complexity but does not really raise new fundamental issues. For example, the Wool data consists of three factors:

Table 4: Regression summary for UN11 data: the main effects model

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   49.529      3.400  14.569  < 2e-16 ***
groupother    -1.535      1.174  -1.308    0.193
groupafrica  -12.170      1.557  -7.814 3.35e-13 ***
log(ppgdp)     3.177      0.316  10.056  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 5.109 on 195 degrees of freedom
Multiple R-squared:  0.7492,Adjusted R-squared:  0.7453
F-statistic: 194.1 on 3 and 195 DF,  p-value: < 2.2e-16
```

Table 5: The Wool data

Variable	Definition
len	Length of test specimen (250, 300, 350 mm)
amp	Amplitude of loading cycle (8, 9, 10 mm)
load	Load put on the specimen (40, 45, 50g)
log(cycles)	Logarithm of the number of cycles until the specimen fails

Different models can be considered for the data:

$$log(cycles) \sim len + amp + load,$$

$$log(cycles) \sim len + amp + load + len : amp$$

$$+ len : load + amp : load,$$

$$log(cycles) \sim len + amp + load + len : amp$$

$$+ len : load + amp : load + len : amp : load.$$

Mean function with only factors and interactions are often called *analysis of variance models*. These models are discussed more completely in experimental design books. Analysis of variance models are really a subset of multiple linear regression models.

3 Polynomial Regression

If a mean function with one predictor X is smooth but not straight, integer powers of the predictors can be used to approximate $E(Y|X)$. The simplest example of this is quadratic regression:

$$E(Y|X = x) = \beta_0 + \beta_1 x + \beta_2 x^2.$$

A more general form is the so-called polynomial regression:

$$E(Y|X = x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_d x^d.$$

3.1 Polynomials with several predictors

With more than one predictor, we can consider the second-order mean function:

$$E(Y|X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2.$$

With k predictors, the second-order model will include $k(k-1)/2$ interaction terms, which can be huge for a large value of k.

3.2 Numerical Issues with Polynomials

Numerical problems can arise when using polynomial regressors in a regression. The first problem is that regressors X^d and X^{d+1} can be very highly correlated, and high correlations can cause inaccurate computation of the OLS estimator. A second problem is that for some problems, X^d can be so large (or, if $|X| < 1$, so small) that significant round-off error occurs.

A solution to these computing problems is to use orthogonal polynomials to define the polynomial regressors. For example, for fitting a cubic polynomial with regressors X, X^2 and X^3, we would fit with regressors $Q_1 = X - \bar{X}$, the residues Q_2 from the regression of X^2 on Q_1, and the residuals Q_3 from the regression of X^3 on Q_1 and Q_2. The Qs are then rescaled to have unit length. The resulting Q_j are uncorrelated, and so replacing

(X, X^2, X^3) by the rescaled Q_j's avoids numerical problems.

4 Splines

A polynomial fit is really just a weighted sum of basis function:

$$E(Y|X = x) = \beta_0 + \sum_{j=1}^{d} \beta_j x^j,$$

where the basis functions are the monomials $\{x^1, x^2, \ldots, x^d\}$.

Splines provide a different set of basis functions, each of which acts locally, so changing the weight for one of the basis functions will mostly affect the fitted curve only for a limited range.

5 Principal Components

Suppose we have variables $X_1, \ldots, X_k$ with k large. Our goal is to replace the k variables with $k_0 < k$ linear combinations of them such that the smaller set of variables represents the larger set as closely as possible. Let start with $k_0 = 1$. Let $\boldsymbol{X}' = (X_1, \ldots, X_k)$ be the variables written as a vector, and let $\boldsymbol{u}_1$ be a $k \times 1$ vector of constants, subject to the constraint that $\boldsymbol{u}_1'\boldsymbol{u}_1 = 1$. The first

principal component will be a linear combination $Z_1 = \boldsymbol{u}'\boldsymbol{X}$ such that the variance of Z_1,

$$\text{Var}(Z_1) = \text{Var}(\boldsymbol{u}_1'\boldsymbol{X}) = \boldsymbol{u}_1'\text{Var}(\boldsymbol{X})\boldsymbol{u}_1,$$

is as large as possible to retain as much as the variation in the predictors as possible. The solution is to set $\boldsymbol{u}_1$ to be the eigenvector corresponding to the largest eigenvalue of $\text{Var}(\boldsymbol{X})$. For a solution with k_0 principal components, the linear combinations are the eigenvectors corresponding to the k_0

largest eigenvalues.

In the usual case, Var$(\boldsymbol{X})$ is unknown, and the sample covariance matrix is used in place of the unknown variance matrix.

Consider the example of “Professor Ratings”.

In this example, about 78% of the variance in the five ratings is captured by the first PC, and about 98% of the variance is captured by the first three PCs.

Table 6: PCA for the example of “Professor Ratings”.

```
eigen() decomposition
$values =variance
[1] 2.393079 0.386789 0.220536 0.056929 0.001344
$cumulative proportions
      0.78       0.91        0.98         1.00        1.00
$vectors
          [,1]       [,2]       [,3]       [,4]       [,5]
[1,] -0.5354480 -0.1549957 -0.1502569 -0.046226  0.8152041
[2,] -0.5291273 -0.1356586 -0.1362543 -0.700748 -0.4381879
[3,] -0.5365282 -0.1882817 -0.1671211  0.710877 -0.3786976
[4,] -0.3357585  0.9164092  0.2145889  0.037289 -0.0046301
[5,] -0.1808964 -0.2869097  0.9406805  0.008903  0.0005212
```

The eigenvector $\hat{u}_1$ gives almost equal weight to the first three scales, and lower weight to the remaining scales. The eigenvector $\hat{u}_2$ is essentially for the rating ”easiness” (with a big weight for it), and $\hat{u}_3$ is essentially for “raterInterest”.

5.1 Using Principal Components

PCs are sometimes used in regression problems to replace several variables by just a few linear

combinations of them. For this example, we might use $Z_j = \boldsymbol{X}'\hat{u}_j$ for $j = 1, 2, 3$ as regressors in the model. In this particular example, we might choose to use the three regressors consisting the average of the first three ratings, easiness, and raterInterest, because they are much easier to interpret.

6 Missing Data

In many problems, some variables will be unrecorded for some cases. The methods we study in this course generally assume and require complete data, without any missing values. The literature on missing data problems is very large, and our goal is more to point out the issues than to provide solutions.

An alternative to deleting cases with missing values is to “fill in” the missing data with plausible values. A solution to this is using *multiple imputation*, in which several filled in data sets are created a complete data analysis is performed for each data set, the results are averaged to get an overall analysis.